肿瘤类疾病的过度与错误医疗检查控制机制与模型的研究 *

朱诗生 [1]，汪昕蓉 [1]，毛礼厅 [2]，柳学国 [2]

(1. 汕头大学 计算机系，广东 汕头 515063; 2. 中山大学附属第五医院，广东 珠海 519000)

摘 要：针对当前肿瘤类疾病诊治过程中存在的错误与过度医疗问题，基于医疗大数据提取出相似病案专家处方中的影像信息，利用机器学习分类模型提出了发现错误与过度诊治的检查控制机制与解决方案。该方案依托医院长期积累的各类肿瘤疾病病历中的 CT、MRI 图像，以每次诊疗过程中的实际肿瘤类型为依据，从医疗数据库中选择对应类型的影像数据进行特征提取、特征选择、模型构建，得到该类型肿瘤的预测分类器，预测当前病例的良恶性；并通过跟医生诊断结果的对比判断诊疗过程中是否存在过度与错误医疗问题。其核心是提高不依赖人工判别方法的判别正确率来降低肿瘤类疾病的错诊可能性，通过实验证明结合了 spearman 去冗余方法的 SVM_RFE 降维，与传统的 SVM_RFE 方法相比，在肺结节良恶性分类问题的 SVM 模型中表现更佳，同时也优于传统的 radiomics 方法。该方案能及时发现错误与过度医疗问题并提出预警，发挥监督提醒的作用，在实现预防和避免诊治错误的同时减少对人工鉴别的依赖，为错误医疗问题及减轻患者负担提供一种新的解决途径。

关键词：错误医疗；机器学习；spearman；SVM_RFE；SVM 分类模型

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2018.04.0267

Study on evaluation mechanism of excessive treatment and misdiagnosis of tumor diseases

Zhu Shisheng[1], Wang Xinrong[1], Mao Liting[2], Liu Xueguo[2]

(1. *Dept. of Computer, Shantou University, Shantou Guangdong* 515063, *China*; 2. *the* 5*th Affiliated Hospitol of Sun Yat-Sen University, Zhuhai Guangdong* 519000, *China*)

Abstract: Aimed at solving the problem of erroneous and excessive medical treatment of tumor diseases, this article extracted image information from similar medical record experts' prescriptions based on medical big data, and used a machine learning classification model to quantitatively analyze the level of medical treatment. The program relies on the CT, MRI images of tumor diseases accumulated in the hospital over a long period of time. Based on the tumor type in each treatment case, it selected the corresponding type of image data from the medical database for feature extraction and feature selection, to construct models to obtain a predictor for this tumor disease, to predict the benign and malignant of the current case, and to determine whether there are excessive and erroneous medical problems in the diagnosis process by comparing with the results of the doctor's diagnosis. The core of the method is to improve the accuracy of discriminating tumors without relying on human beings. Compared with the traditional SVM_RFE, the investigated method which combines SVM_RFE with the Spearman correlation has been experimentally proven to perform better in the SVM model of benign and malignant classification of pulmonary nodules. In general, it offers better performance compared with traditional radiomics method. The solution can detect erroneous and excessive medical treatment issues in real-time and provide warning, which can play a role in supervision and reminding. It potentially reduces the reliance on manual identification and minimizes the burden on patients while preventing and avoiding errors in diagnosis and treatment.

Key words: error treatment; machine learning; spearman; SVM_RFE; SVM classifier

0 引言

近年来，人们的生活环境发生了极大的变化，同时癌症的发病率也在不断上升。根据中国医学科学院肿瘤医院赫捷院士、全国肿瘤登记中心主任陈万清教授等人在2016年1月发表的文章《Cancer statistics in China, 2015》显示，中国作为拥有 13.7 亿人的人口大国，在 2015 年一年时间内，预计有 429.2 万新发肿瘤病例和 281.4 万死亡病例。这其中肺癌的发病和死亡率居

收稿日期：2018-04-07；**修回日期：**2018-05-17　　**基金项目：**广东省科技计划资助项目（20140401）

作者简介：朱诗生（1963-），男，广东汕头人，副教授，硕士，主要研究方向为医疗流程控制、移动医疗（sszhu@stu.edu.cn）；汪昕蓉（1991-），女，广西桂林人，硕士研究生，主要研究方向为医疗流程控制、医学图像处理、数据挖掘；毛礼厅（1990-），女，江西上饶人，住院医师，硕士，主要研究方向为胸腹部影像诊断；柳学国（1964-），男，湖北武人，主任医师，博士，主要研究方向为胸腹部影像诊断.

各类癌症之首，平均每 1.2 min 就有一个中国人死于肺癌[1]。随着科学的进步和治疗手段的发展，人类面对各种疾病时有了更多的治疗选择，但是，各种药物和新型疗法层出不穷，医疗诊治过程中的错误与过度医疗问题日益严重。错误与过度医疗是指在诊治过程中，由于医生主观或疏忽而导致患者病情更加严重或加重患者负担的情形，这其中包括但不限于错误诊断、错误用药、手术失误等[2]。

肿瘤疾病死亡率高，治疗开销大，其错误诊治将给病人带来极大的痛苦和负担。目前临床上对肿瘤的良恶性的初步诊断主要依靠医生的主观经验与某些临床特征的相结合，在医生诊断怀疑是恶性肿瘤之后才会做进一步确认。对于良性肿瘤病人而言，被误判为恶性肿瘤所做的一系列复检和治疗实际上是过度医疗；而被误判为良性的恶性肿瘤病人则可能错失尽早治疗的良机。因而解决肿瘤的良恶性分类问题成为减少肿瘤类疾病的错误与过度医疗问题的关键。

错误医疗由于其偶发性及评判标准的模糊性，长期以来该领域的研究并没有太多突破性的进展。一些国家主要依靠道德约束和舆论压力减少错误与过度医疗现象，或者是通过规范化用药流程，将药品的使用置于专业医疗人员的监督控制下，来预防和避免错误用药和滥用药品问题的发生，并未能从技术手段上实现对错误与过度医疗的监控与管理[3]。2012 年 Lambin 教授提出的 radiomics 影像组学理念在医疗科研工作者中掀起一股基于医疗图像预测疾病类型的热潮。接下来的几年，对于肿瘤类重大疾病，一种基于图像分析结合统计学线性回归方法来预测疾病程度的研究被许多研究者不断推进，旨在改变传统医生诊断肿瘤良恶性只能依靠读片经验的现状[4]。此类预测模型的构建正是肿瘤类疾病过度与错误医疗问题解决方案的突破口，2016 年学者 Carneiro 通过基于 CT 图像对老年人 5 年内死亡率的预测，展示了机器学习的非线性分类方法在疾病预测方向上的潜力和优势[5]，为错误医疗问题的解决提供了另一种思路。

本文的研究重点是基于肿瘤类病人病历构建可信度高的良恶性分类模型，降低过度与错误医疗的发生概率。对于收集的 194 例肺结节病例，文中先使用当前比较主流的 radiomics 方法：从病历包含的医学图片中提取 Texture 特征、lasso regression 降维、logistics regression 建模，使用 ROC 曲线对该模型进行评估；再在降维和建模的环节中采用其他统计和机器学习方法，得出不同的特征选择方法与建模方案在该问题上的综合比较。最终得到引入 spearman 去冗余方法后，SVM_RFE 降维和 SVM 分类器结合的模型在肺结节良恶性鉴别的问题上表现最佳的结论。

1 过度与错误医疗问题发现机制的框架研究

本研究在整理大量肿瘤类疾病的医疗数据，深入调查理解医院医疗流程的基础上，提出了针对肿瘤类疾病错误与过度诊治的检查控制机制模型与方法；该机制将从病人的病历入手，使用病历中的影像信息，通过机器学习方法做出肿瘤的良恶性预测，与医生的实际诊断做对比形成评估报告，从而实现对可能出现的错误诊治、过度医疗现象做出预警监督；对其中预测结果与实际诊断差异较大，出现错误诊治率高于阈值的评估结果提交上级监管部门，帮助医生及时发现潜在的错误与过度医疗问题，避免该类问题的发生。该方案的目的是通过基于影像组学的方法对大量肿瘤类病症进行数据挖掘，对诊治过程是否存在过度与错误医疗问题提出一种科学规范的评判解决机制。旨在提高评判标准的客观性和合理性，减少对人工鉴别的依赖，具体的监控流程框架如图 1 所示。

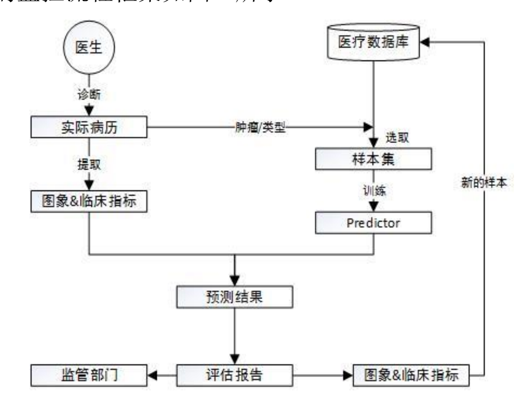


图 1　过度与错误诊治问题发现机制与方法的流程框架

从上图展示的错误与过度医疗问题发现机制与方法流程框架可看出，该框架主要是将医生参与程度高的诊治流程与诊治结果的评价流程分离开来，尽量降低评估报告中人为因素的影响，提高最后得到的评估报告的客观性。通过基于医疗数据库中大量已有病历的影像数据，提高预测评估模型的可重复性和可信度。肿瘤病人到医院就诊，医生将结合病人对病症的描述，以及相关的一系列医疗检查，最后根据检查结果和个人的知识经验作出诊断，与其他信息一起形成了该次诊治的实际病历，病历中就包含了 CT 图像、MRI 图像等影像数据。

该病历的具体肿瘤类型作为过滤器的输入，通过过滤器对医疗数据库中的多种肿瘤影像数据进行有效筛选、特征提取，用于下一步的机器学习，目的是建立与实际诊断结果相对应类型肿瘤的预测分类模型；而病历中的影像数据部分，之后则作为该预测模型的输入样本，提取出图像多类特征属性的离散或连续值，放入训练好的预测模型 predictor 中进行预测分类，根据分类结果与输入端的临床良恶性指标即实际诊断结果的异同，得到此次诊治过程是否存在错误医疗现象的预判。实际诊断和预测结果不同将被判定为可能存在错误医疗的诊治流程，并会在预警评估报告中指出错误诊治存在的可能性。二者相同则该病历被判定不存在错误和过度医疗情况，图 1 的过度医疗发现机制将认为此次诊治流程是规范的，该病历中的影像数据和诊断结果应该作为可信赖的新样本添加入医疗数据库中，医疗数据库中除了保留该病历的影像数据之外，还将保留该次诊断结果即临床良恶性指标作为其对应的建模标签。

如图 1 所示，医疗数据库是预测模型建立的基础，其中除存储病理、药物知识，还包括病案专家处方库，数据库需要不

断积累，此环节因篇幅不做展开。实际诊治中各类型疾病所对应的检测项目不同，因此得到的病理参数也不一致。例如可能患有肿瘤的病人需要做 CT 或 MRI 扫描，得到的病理数据为非结构化的图像数据；而疑似病毒性感冒的患者则需要做常规的血液检查，得到的病理数据则是结构化的文本数据，本研究项目组开发的协同诊治平台对此已做了方案设计[6]。本文针对的疾病类型是肿瘤，因此只考虑医疗数据库中的肿瘤类疾病数据的积累。在医疗数据库建立的初期，为了保证所建立模型的正确性，需要以绝对可信的样本为初始训练集，因此一开始数据库中保存的肿瘤影像数据必须来源于最终做了手术切除的病例，以肿瘤切除后病理证实的结果作为影像的对应金指标。

2　过度与错误医疗问题建模方法研究

2.1　过度与错误医疗问题分类模型

错误与过度医疗问题解决的重点在于建立可靠的分类模型，来判定医生的实际诊断中是否存在诊治错误。支持向量机，英文名为 support vector machine，简称 SVM，是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解，契合过度和错误医疗问题对肿瘤良恶性分类的需求，本文将把该理论模型用于解决错误医疗的问题。

依据文献[7]中所述支持向量机原理，给定某种肿瘤疾病的良性样本和该类肿瘤的恶性样本作为总体训练样本集 D={(x_1，y_1)，(x_2，y_2),…, (x_m，y_m)}，其中 x_i 是由病历中影像数据提取的各类图像特征参数，y_i 表示该病历所对应患者的真实肿瘤良恶性级别，分类学习最基本的想法就是基于训练集 D 在样本空间中找到一个划分超平面，将不同类别的诊治样本分开。标记分类目标疾病为 y_1，其他类型疾病 y2……ym 统一标记为 y0，我们的理想目标是寻找位于 y0 和 y1 两类训练样本“正中间”的划分超平面，在样本空间中，划分超平面可通过如下线性方程来描述[7]：

$$\omega^{\mathrm{T}} x + b = 0 \tag{1}$$

其中：$\omega = (\omega_1;\omega_2;\omega_3;\ldots;\omega_d;)$ 为法向量，决定了超平面的方向，b 为位移项，决定了超平面与原点之间的距离。显然划分超平面可被法向量 ω 与位移 b 确定，下面我们将其记为 (ω,b)。样本空间中任意点到超平面 (ω,b) 的距离可写为

$$\mathrm{r} = \frac{\left|\omega^{\mathrm{T}} x + b\right|}{\|\omega\|} \tag{2}$$

假设超平面 (ω,b) 能将 y0 和 y1 两类训练样本正确分类，即对于 $(x_i, y_i) \in D$，y_i 若属于分类目标疾病 y1，则有 $\omega^{\mathrm{T}} x + b > 0$；若不属于分类目标疾病 y1，则有 $\omega^{\mathrm{T}} \chi + b < 0$。令

$$\begin{cases} \omega^{\mathrm{T}} x + b \geq +1, & y_i = y_1; \\ \omega^{\mathrm{T}} x + b \leq -1, & y_i = y_0; \end{cases} \tag{3}$$

使得上式成立的训练样本点，即对应的病历影像数据，它们被称为“支持向量”，两个异类支持向量到超平面的距离之和为

$$\gamma = \frac{2}{\|\omega\|} \tag{4}$$

它被称为“间隔”[7]。

所需过度与错误医疗问题分类模型的工作流程为根据病历中所描述的肿瘤疾病类型结合其对应的影像数据构建分类器。CT 或 MRI 图像的采集目的在于尽可能正确的识别出恶性的肿瘤病例，因此我们在医疗数据库中挑选一定数量的某类肿瘤恶性病例的影像数据作为 SVM 训练的正样本集，另选取数量相等的该类肿瘤良性病例的影像数作为训练集中的负样本集。通过对这些数据的训练，寻找到最优划分超平面，即获得最终的预测模型，用于鉴别肿瘤的良恶性，并以此预测结果和医生诊断的异同作为是否存在过度与错误医疗的评判依据。

2.2　疾病分类特征选择算法

考虑到某些重大疾病获得的临床特征较多，特别是采集影像数据的肿瘤类疾病，可获取的基于图像的特征信息量巨大，在使用 SVM 建立预测模型时容易因为训练样本的数据维度过高产生过拟合现象，最终影响分类模型的泛化能力，因此对训练数据进行特征选择至关重要。临床信息类别较多较复杂的疾病，特征选择能够减少待处理的数据量，降低计算复杂度，缩短建立模型的时间消耗；同时，并不是所有的特征属性都能对预测模型起到正面作用，特征选择能够降低数据维度，将具有干扰性的特征筛选出来，从而提高错误医疗分类模型的预测准确率。SVM_RFE 特征选择算法是一种以支持向量机为基础的特征排序选择算法[8]。该算法的实现分为以下三个步骤：

a)使用包含多个临床特征参数的待训练数据集训练支持向量机，获得 SVM 正则项的拉格朗日算子向量，即分类器参数；

$$\alpha = SVM - train(X, y) \tag{5}$$

b)计算每个特征向量即临床指标参数在输入空间中的权重向量，得到完整支持向量机的权重；

$$\omega = \sum_k \alpha_k y_k X_k \tag{6}$$

c)使用合适的排序方法将数据集里的权重向量进行排序并删除贡献率最小的特征，这说明该项指标在正确分类目标疾病的过程中作用最低，直到最后数据集中只剩下一个特征。

$$R_c = \left|\omega^2 - \omega^{-(p)^2}\right| \tag{7}$$

ω^2 表示的是整个支持向量机的权重，$\omega^{-(p)^2}$ 则表示剔除了第 p 个特征之后支持向量机的权重[9]。

这种特征排序的方法基于 SVM 在训练过程中产生的特征权重，选出来的参数与分类结果相关性大，因此与 SVM 分类器配合使用有较好的预测效果。但显然在上述的特征排序过程中，使用的是特征逐次减少的方法，并不能充分考虑特征之间的相互联系，可能出现某几个相关程度高的特征同时排在序列前段，挤压特征选择空间，从而不能找到最优特征组合的情况。对此，本文采取在 SVM_RFE 特征选择算法之前先使用 spearman 基于相关性去冗余的方法，删除特征中相关程度高的部分特征。

3　结合 spearman 的 SVM 方法与传统 radiomics 方法的异同

3.1　训练数据的提取流程

前文介绍了基于 SVM 的错误医疗问题建模方法，接下来要解决的问题就是获得高质量的训练数据，以保证最后获得预测模型的泛化能力。上文的框架描述中已提到模型的训练数据来源于医院数据库中已有肿瘤病历的影像数据，具体流程如图 2 所示。

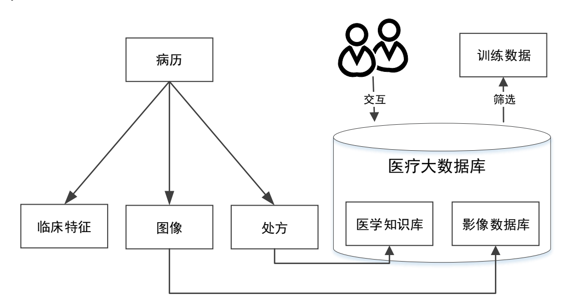


图 2　基于相似专家处方的训练数据提取流程

图 2 是基于医疗数据库中大量已有病历提取训练数据集的模型；其中病理、药物知识库用于存储从医学专家那里得到的针对某类疾病的专门知识和经验，以及医院长年累月积累下来的优秀病历及处方，数据库的丰富性和数据正确性直接关系到最终预测模型的可信度，数据的表现形式为规范化的病历数据，其内容主要有：病人信息、病症描述、病理参数、医生诊断医嘱[10]，以及非结构化的影像数据。预测模型的构建是以待评估病人的实际患病类型为依据，通过人机交互从医疗数据库中提取出一定数量该类肿瘤恶性病例的影像数据作为训练集的正样本，并取出数量相等的该类肿瘤良性病例影像数据作为负样本；对这些从数据库中随机挑选出来的待训练样本集进行特征提取，通过结合了 spearman 去冗余方法的组合 SVM-RFE 特征选择算法进行数据降维，降维后的结果用于构建 SVM 分类器，因此，医疗数据库中训练数据的规范收集和积累是分类器预测效果的保障。

3.2　基于传统 radiomics 方法的肿瘤分类模型构建

如前文所述，错误医疗问题的解决关键在于依据病理指标正确地将疾病分类，就肿瘤论，若能有效降低肿瘤类重大型疾病在诊治过程的错误医疗问题，对患者意义重大，既能减少无谓的花费，又能及时接受合适的治疗。当前影像组学一大研究方向是从肿瘤影像数据出发，依靠计算机领域的图像处理技术和统计学领域的统计分析、线性回归等方法构造出肿瘤疾病的分级和分类模型，将肿瘤疾病的预测与临床医学经验知识的积累相剥离开，评估这种预测方案的准确性和 AUC 等指标。近些年来 radiomics 领域较为流行的方法，主要分为如图 3 所示的四个步骤：a)数据预处理，包括 CT/MRI 图像的采集，专业医生对感兴趣区域（一般是病灶区）的勾画，和基于灰度值图像的特征提取；b)为了获得线性模型的稀疏化可行解，使用 lasso regression 对上一步提取的特征做降维；c)基于步骤 b)中挑选出来的特征参数，使用 logistic regression 构建肿瘤良恶性分类模型；d)使用留出的验证集部分数据对步骤三所构造的模型进行评估，采用包括预测的准确率、预测样本的灵敏度和特异度等多重指标对该模型的泛化能力进行总体评价 [11]。

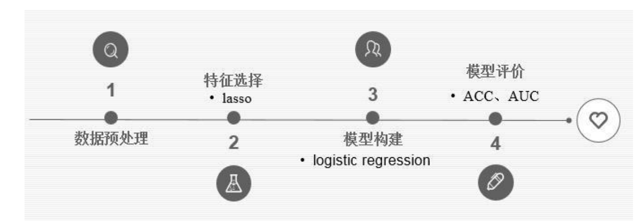


图 3　基于 radiomics 方法的肿瘤分类模型

3.3　基于 SVM 和相似专家处方的疾病预测模型构建

本文所研究疾病肺结节的 CT 影像数据如图 4 所示，图 4 中左圈出位置可见病焰区域占比很小，图右是放大多倍以后的结节 3D 图像，医生单凭这些影像信息判断结节的良恶性难度很大，容易导致错误医疗情况的发生。

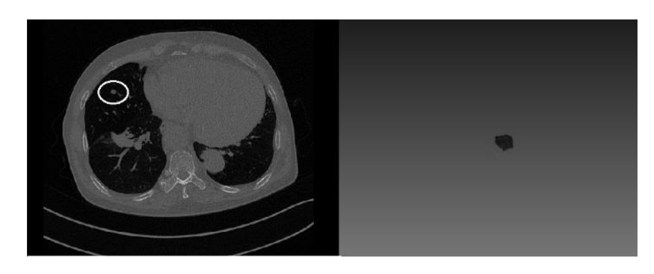


图 4　实验所用肺结节 CT 图像

由于人体各个组织的病灶在医学图像上有不同的表现，相似病例的临床特征尤其是影像数据搜集较为困难，这也使得可用于模型训练的样本量不够大，更适合用小样本分类表现良好的机器学习方法。本研究建模流程为：a)收集处方中已得出良恶性病理标签的肿瘤 CT 图像数据 N 例，将其随机划分成训练集数据 x 例和验证集数据 y 例；b)全部样本通过图像识别的方法提取出 GLCM、GLRLM、Histogram、Form factor 四类纹理参数共 M 个，得到 N*M 的特征值矩阵；c)基于训练集样本进行特征工程，结合 spearman 方法筛选出对模型分类最有效的部分特征 m 个，特征值矩阵降维到 N*m，降维后的部分数据如图 5 所示；d)P 基于 x 例训练样本构建 SVM 分类模型，用 y 例验证集样本对该模型进行验证，评估模型的疾病分类预测能力。

A	B	C	D	E	F	G	H	I	J
label	MeanValue	stdDeviatic	Variance	VolumeCo	VoxelValue	Range	RMS	MeanDevi	RelativeDe
0	0.004712	35421.1	0.000135	0.000155	1.89E+08	35421.1	0.997243	35421.1	21.6536
1	0.017474	27815.8	0.000424	0.00102	3.3E+08	27815.8	0.989583	27815.8	99.3137
1	0.034131	18932.6	0.001794	0.003115	3.52E+08	18932.6	0.980263	18932.7	162.118
1	0.026458	19031.7	0	0.001776	3.8E+08	19031.7	0.984375	19031.8	0
1	0.015748	15001.2	0.000634	0.001363	53500000	15001.2	0.991071	15001.2	57.7937
1	0.00869	22188.5	0.000196	0.000369	1.16E+08	22188.5	0.994898	22188.5	103.561
0	0.002451	30315.4	7.91E-06	2.80E-05	1.52E+08	30315.4	0.999232	30297.8	16.6707
1	0.004915	23587.6	0.000237	0.00013	1.43E+08	23587.6	0.99717	23587.6	67.6166
1	0.007647	24242.9	0.000132	0.000169	1.09E+08	24242.9	0.995455	24242.9	60.3299
0	0.000909	33702.3	7.56E-06	1.03E-05	87200000	33712.7	0.999483	33695.1	35.7516
0	0.000897	29989.1	1.29E-05	1.60E-05	16100000	29993.8	0.999486	29991.3	20.7221
1	0.006067	25732	8.06E-05	0.000246	1.87E+08	25732	0.996445	25732.1	82.2378
0	0.001736	34417.3	3.67E-05	2.53E-05	90900000	34417.3	0.998996	34399.2	13.8518
1	0.010823	13908.9	0.000299	0.000643	87300000	13908.9	0.993697	13908.9	43.1621
1	0.018699	31257.2	0.00049	0.00147	1.32E+08	31257.2	0.988806	31257.2	39.3108
0	0.005133	25338.6	6.77E-05	0.00011	1.51E+08	25338.6	0.996988	25338.6	63.1897

图 5　降维之后的训练数据集

3.4　结合 spearman 的 SVM 方法与传统 radiomics 方法的比较

由 3.1 与 3.2 节对传统 radiomics 方法与结合了 spearman 的 SVM 方法的步骤说明可知，两种方法作用于肿瘤疾病过度与错误医疗检查控制机制的方式是相同的，都是从数据库中获取影像数据，反馈良恶性预测结果。且两种方法在数据的预处理、特征提取以及最后模型评估环节基本相同，不同之处在于特征降维和分类器构建。Radiomics 方法构建的是一般的线性回归模型：logistic regression，并在特种工程环节为了获得稀疏解采用了 lasso regression：带 L1 约束的线性回归模型，lasso 本身不具备相关性去冗余功能；本文使用的结合了 spearman 的 SVM 方法在建模环节采用了带高斯核的 SVM 分类器，一种非线性的分类器，在线性可分程度不高的数据集上也能有良好表现，并在特征降维环节结合使用 spearman 相关性去冗余方法，通过 SVM_RFE 非线性的降维排序法更大程度的保留了有效特征。

4　实验验证及结果说明

本文采用中山大学第五附属医院收集的 194 例真实肺结节影像数据作为实验样本，通过实验与一般的 SVM 方法和传统的 radiomics 方法两两比较，并且对实验结果进行了可视化呈现，一定程度上验证了本文所提出的检查控制机制与模型的合理性和科学性。当然，该方案在临床应用之前还需要做更多的适用性研究。

本次验证实验的主要过程分为三个步骤，分别是对数据的预处理、特征降维与建模、模型的评估与比较。用验证集样本分别对 3 个模型做评估预测后，最终结合了 spearman 去冗余的改进 SVM 模型在 59 例验证数据中表现最佳，达到了 89.83% 的分类预测准确率。

4.1　实验数据划分及预处理

实验总样本为肺结节病例数据 194 例，其中良性结节 139 例，恶性结节 55 例，实验采用常用的 7：3 作为训练集与验证集的划分比例。按照正负样本比例随机选取 135 例用于做训练，训练集中良性结节 98 个，恶性结节 37 个；59 例用于做模型验证，验证集中良性结节 41 个，恶性结节 18 个；对上述收集的样本通过图像处理方法获取 GLCM、GLRLM、Histogram、Form factor 四类纹理特征共 330 个，篇幅原因本文不细述特征的提取过程。接着分别对提取出来的训练集和验证集数据进行标准化和异常值均值填充处理，赋予恶性肿瘤数据 0 类标签，良性肿瘤的标签设置为 1。数据划分和预处理之后所得样本的具体分布如表 1 所示。

表 1　实验数据划分结果

标签	0 类	1 类	总计/例
训练数据	37	98	135
验证数据	18	41	59

4.2　特征降维与建模

本实验的数据降维操作仅使用训练集数据作为降维数据依据，全过程不涉及到验证集数据，仅在模型评估环节将降维步骤选中的特征在验证集样本中对应挑出，用作每个模型的输入。首先根据传统的 radiomics 方法基于 135 例特征维度为 330 的训练数据，使用 lasso 算法降维，直接获得维度为 10 的稀疏解，并且在阈值选取为 0.795 时获得最大收益，此时训练集准确率为 91.11%。然后基于同样的数据集，重新使用 spearman 方法基于相关性去冗余，在相关度阈值设为 0.85 时，剩余特征个数 49；将训练集的标签及 49 个特征构成的 135*50 特征值矩阵作为输入，使用 SVM_RFE 算法进行降维排序，最后对有序的特征进行逐步回归建模，由于支持向量机的特征增加过程中存在休斯效应[12]，通过对训练集做十折交叉验证发现在特征个数选取为 42 时模型表现达到最佳，此时模型在训练集上的内部验证准确率为 97.78%，故选取该 42 个特征用于最后建模。同理对未经过 spearman 降维处理的 330 个初始特征做 SVM_RFE 降维排序，通过交叉验证发现在特征个数选取为 32 时模型表现达到最佳，此时的内部验证准确率为 97.78%，故选取该 32 个特征用于传统 SVM 建模。具体结果如表 2 所示。

表 2　三种不同降维与建模方法在训练集上的实验结果

降维与建模方法	保留特征个数	准确率
Lasso+LR	10	91.11%
Spr+RFE+SVM	42	97.78%
SVM_RFE+SVM	32	97.78%

4.3　模型评估与比较

首先本实验根据 SVM_RFE 选出的 32 个特征，使用带高斯核的支持向量机对训练集建模获得模型 m0，再将训练集中 spearman+SVM_RFE 选中的 42 特征同样使用高斯核 SVM 建模得到模型 m1，lasso 选出的 10 个特征，使用 logistics regression 对训练集建模获得模型 m2。为了获得拟合能力和泛化能力都足够优秀的模型，分别将训练集和验证集数据依次带入三个模型，通过 ROC 曲线、准确率、AUC、灵敏度和特异度五个指标对模型进行综合评估。图 6 和 7 分别是三个模型在训练集和验证集上的 ROC 曲线，由图中曲线可知三个模型在训练集的分类预测上表现非常接近，并且都表现优秀，因此可以认为三个模型的拟合能力都很强，差异不大。在验证集上三个模型的 ROC 曲线都远离横纵轴对角线，这是泛化能力良好的表现，但三条曲线出现了较明显的差异，未使用 spearman 降维的模型 m0 的曲线被使用了该方法的 m1 包围，说明 m1 模型在验证集上的表现优于 m0；基于 SVM 方法的模型与基于 logistics

regression 的影像组学模型有交点，需要通过更多的指标进行评估。

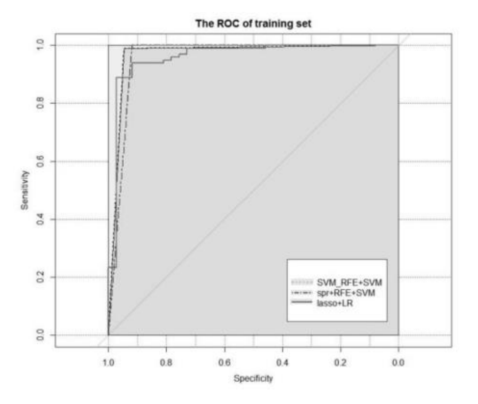


图 6　三种不同方法在训练集上的 ROC 曲线

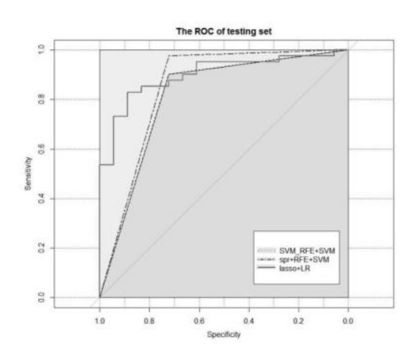


图 7　三种不同方法在验证集上的 ROC 曲线

为了进一步比较三个模型的泛化能力差异，本实验考察了各模型在准确率、AUC、灵敏度和特异度上的表现，详细数据见表 3。由于 SVM 模型构建的超平面到不同类别支持向量是等间隔的，不存在阈值变化，因此得到 ROC 曲线与逻辑回归不同，不存在很多折点，因此难以从 AUC 值上与传统 radiomics 模型做出比较，但在预测准确率和临床灵敏度上有较显著的优势。为了再进一步说明模型 m1 优于其他两个模型，实验最后通过三者的临床决策曲线将各模型的决策净收益进行了可视化，如图 8 所示。

表 3　三个模型在验证集上的评估指标数据

模型	准确率	AUC	灵敏度	特异度
m0	84.75%	0.812	0.902	0.722
m1	89.83%	0.849	0.976	0.722
m2	84.75%	0.900	0.889	0.829

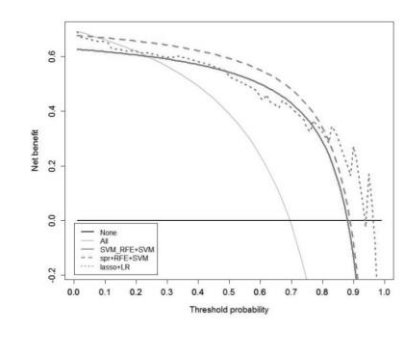


图 8　三种不同方法在验证集上的临床决策曲线

由图 8 中三条决策曲线可以看出，在阈概率 0.2~0.85 区间，模型 m1 的决策净收益要显著高于另外两个模型，并且可选阈概率范围连续并足够大，是良好的判断指标。综上，实验证明结合了 spearman 去冗余算法及 SVM_RFE 降维排序算法的 SVM 模型，能够有效支持本文所提出的肿瘤类疾病过度与错误诊治的检查控制机制的核心评判环节，为降低肿瘤疾病诊治过程中的过度与错误医疗发生概率起到良性监督作用。

参考文献：

[1] Chen Wanqing, Zheng Rongshou, Baade P D, *et al.* Cancer statistics in China, 2015 [J]. CA Cancer J Clin, 2016; 66 (2): 115-32.

[2] Lu Zhifang. On the regretful death of Professor//Dr. Shi Ying-Kang [J]. Quant Imaging Med Surg. , 2016, 6 (3): 334–337.

[3] Kallianidou K, Kiakou M, Tsoukalas N, *et al.* Medication administration in hospital: Difficulties and errors [J]. Arch Hellen Med, 2017, 34 (1): 123-126.

[4] Lambin P, Rios-Velazquez E, Leijenaar R, *et al.* Radiomics: extracting more information from medical images using advanced feature analysis [J]. European Journal of Cancer, 2012, 48 (4): 441-6.

[5] Carneiro G, Oakdenrayner L, Bradley A P, *et al.* Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography [C]// Proc of the 14th IEEE International Symposium on Biomedical Imaging. 2017.

[6] 刘文华. 基于 E-Health 的协同诊治平台的医疗诊治行为检查方案的研究 [D]. 汕头: 汕头大学, 2015. 1-42. (Liu Wenhua. Study on medical treatment behavior check program based on E-health collaborative diagnosis and treatment platform [D]. Shantou: Shantou University, 2015. 1-42.)

[7] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016. (Zhou Zhihua. Machine learning [M]. Beijing: Tsinghua University Press, 2016: 121-125.)

[8] Guyon I, Weston J, Barnhill S, *et al.* Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46 (1): 389-422.

[9] Huang M L, Hung Y H, Lee W M, *et al.* SVM-RFE based feature selection and taguchi parameters optimization for multiclass SVM classifier [J]. Scientific World Journal, 2014, 2014: 795624: 1-10.

[10] 刘洋, 张卓, 周清雷. 医疗健康数据的模糊粗糙集规则挖掘方法研究 [J]. 计算机科学, 2014, 41 (12): 164-167. (Liu Yang, Zhang Zhou, Zhou Qinglei. Study on fuzzy rough set rule mining method for medical health data [J]. Computer Science, 2014, 41 (12): 164-167.)

[11] Huang Y Q, Liang C H, He L, *et al.* Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer [J]. Science Foundation in China, 2016, 34 (4): 2157-2164.

[12] Pal M, Foody G M. Feature selection for classification of hyperspectral data by SVM [J]. IEEE Trans on Geoscience & Remote Sensing, 2010, 48 (5): 2297-2307.